SMART OCR FOR DOCUMENT DIGITIZATION PROJECT REPORT

-D.BALA KOTESWARA SASTRY

1.INTRODUCTION:

A large number of efforts have been put forward that attempts to transform a document image to format understandable for machine so that it can recognize the text or the information from the image. OCR i.e. Optical Character Recognition/Reader provides a solution for this. OCR is software that converts printed text and images into digitized form such that it can be manipulated by machine. OCR systems have established a niche place in pattern recognition. OCR has two categories, online and offline. The image of the scanned document goes through various stages like pre-processing, segmentation, feature extraction, etc. in order to retrieve the information from the image. "*Tesseract*" is one of the most widely used open source library for implementing OCR.

With the advent of OCR techniques, much time has been saved by automatically extracting the text out of a digital image of any invoice or a document. Currently, most organizations that use OCR for any form of automation are for digital copies of invoices or documents are obtained by scanning or taking pictures. The text is extracted from these documents is entered into a template-based data entry software.

1.1Overview:

Text extraction can be achieved by applying text detection that identifies image parts containing text, text localization finds the exact position of the text, text segmentation separates the text from its background and binarization process converts the coloured images into binary. On this binary image, character recognition is applied to convert it into ASCII text. Text extraction is used in creating e-books from scanned books, image searching from a collection of visual data etc...

Necessary Installations

To complete this project you should have the following packages and Softwares

- ☞ Python IDE
- ☞ pytesseract
- ☞ pdf2image
- ☞ tesseract-ocr execution file
- ☞ poppler
- ☞ Flask

1.2PURPOSE

The project aims at creating an application from where the user can upload a pdf document, the document is analysed by Optical character recognition package to extract text from it. The extracted text is again saved in a text document in the local drive.

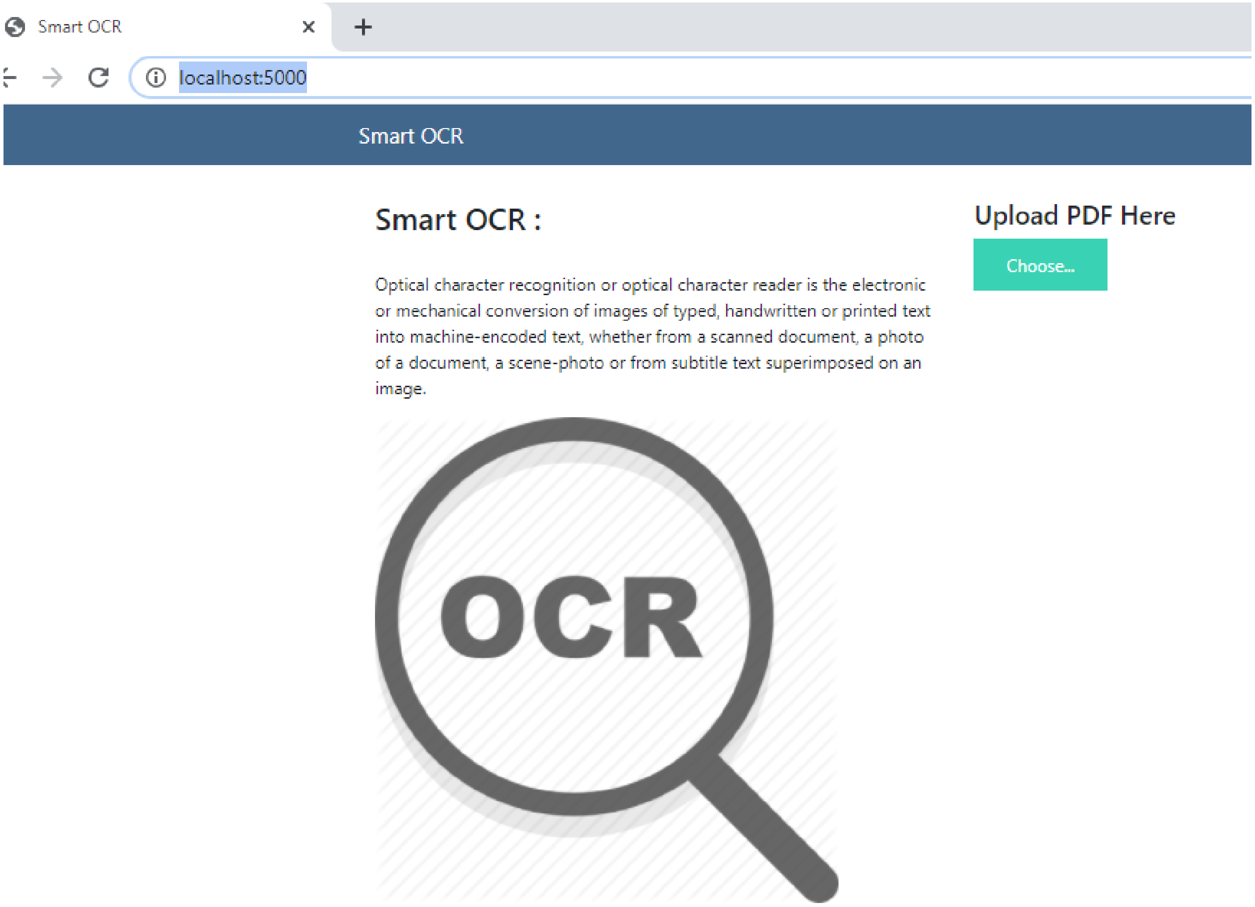
Project Flow:

1. Upload a pdf document
2. Convert PDF document to image
3. Extract the text from the image
4. Store the extracted text in the text document

2.RESULT:

The result obtained by the following steps:

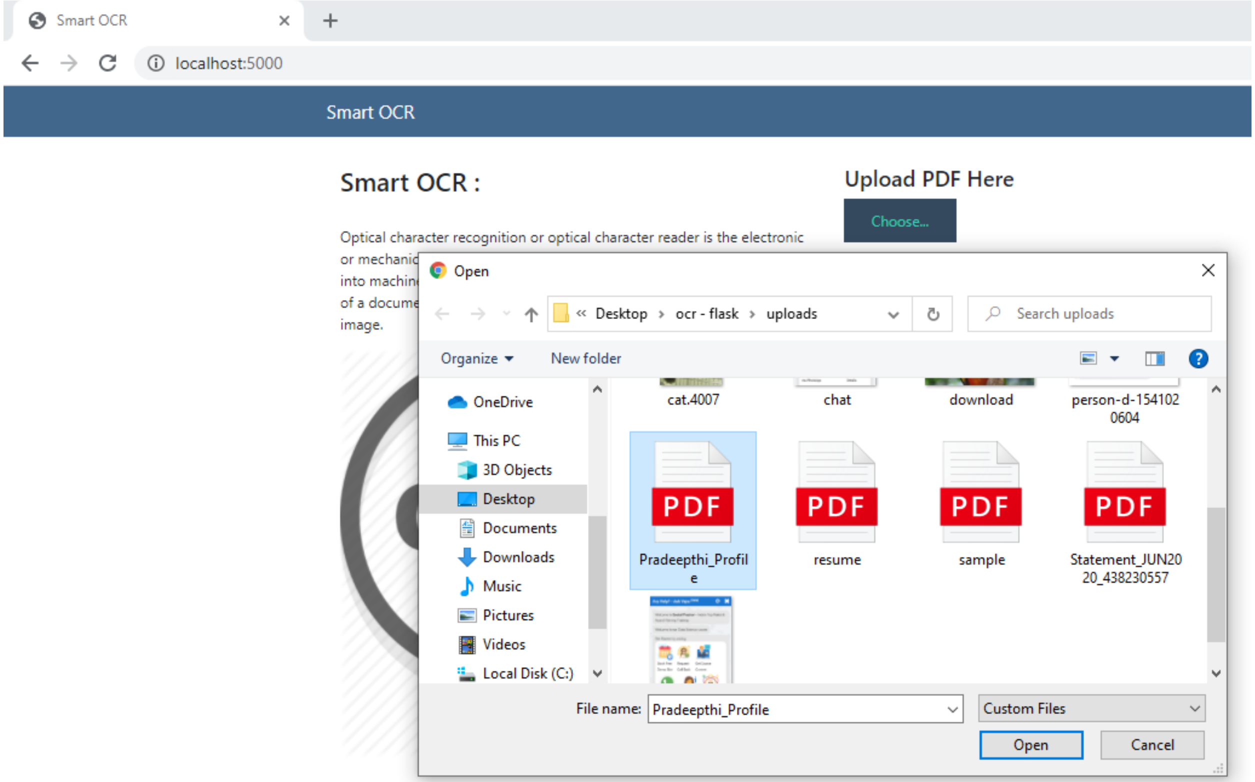
1. Open anaconda prompt/command from the start menu
2. Navigate to the folder where your app.py resides
3. Now type “python hello.py” command
4. It will show the local host where your app is running.
5. Navigate to the localhost where you can view your web page



Smart OCR :

Optical character recognition or optical character reader is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image.

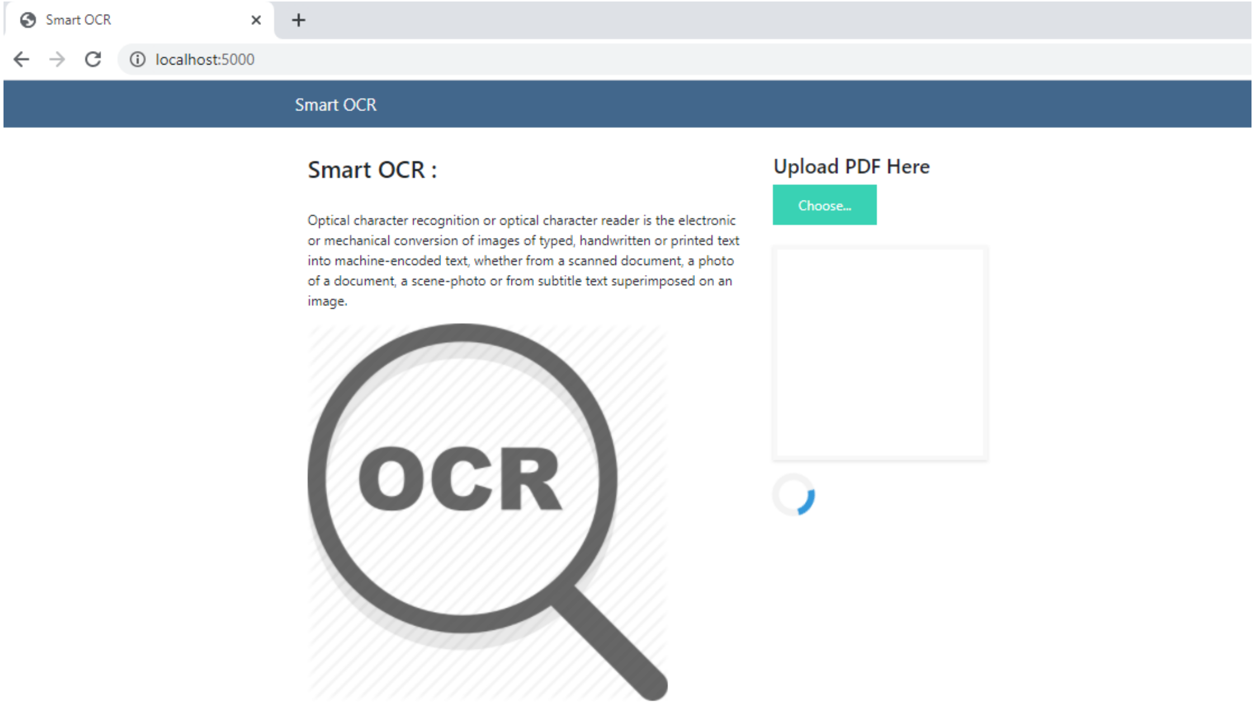
Upload PDF Here



Smart OCR :

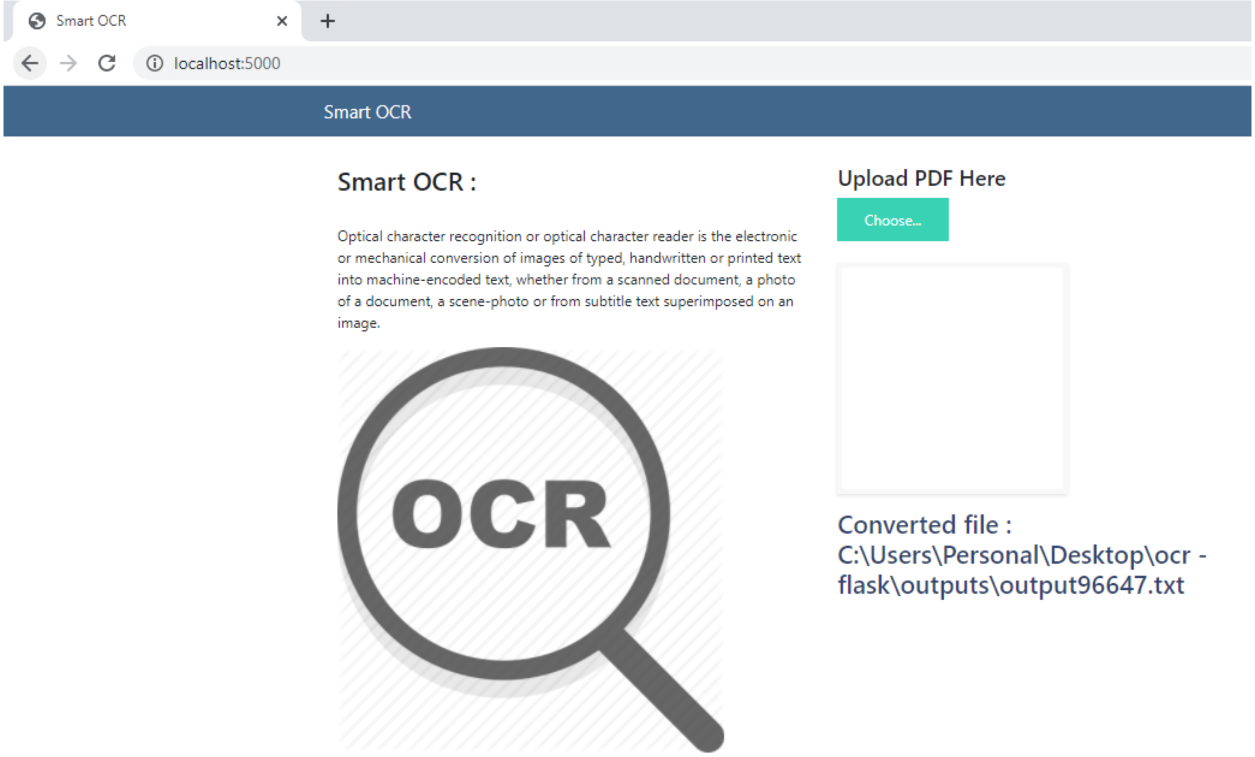
Optical character recognition or optical character reader is the electronic or mechanic into machin of a docume image.

File name: Pradeepthi_Profile



Smart OCR :

Optical character recognition or optical character reader is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image.

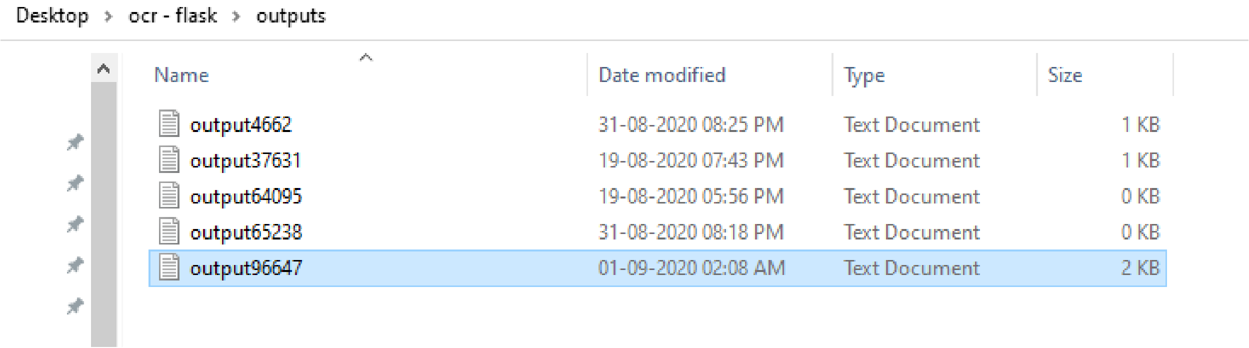


Smart OCR :

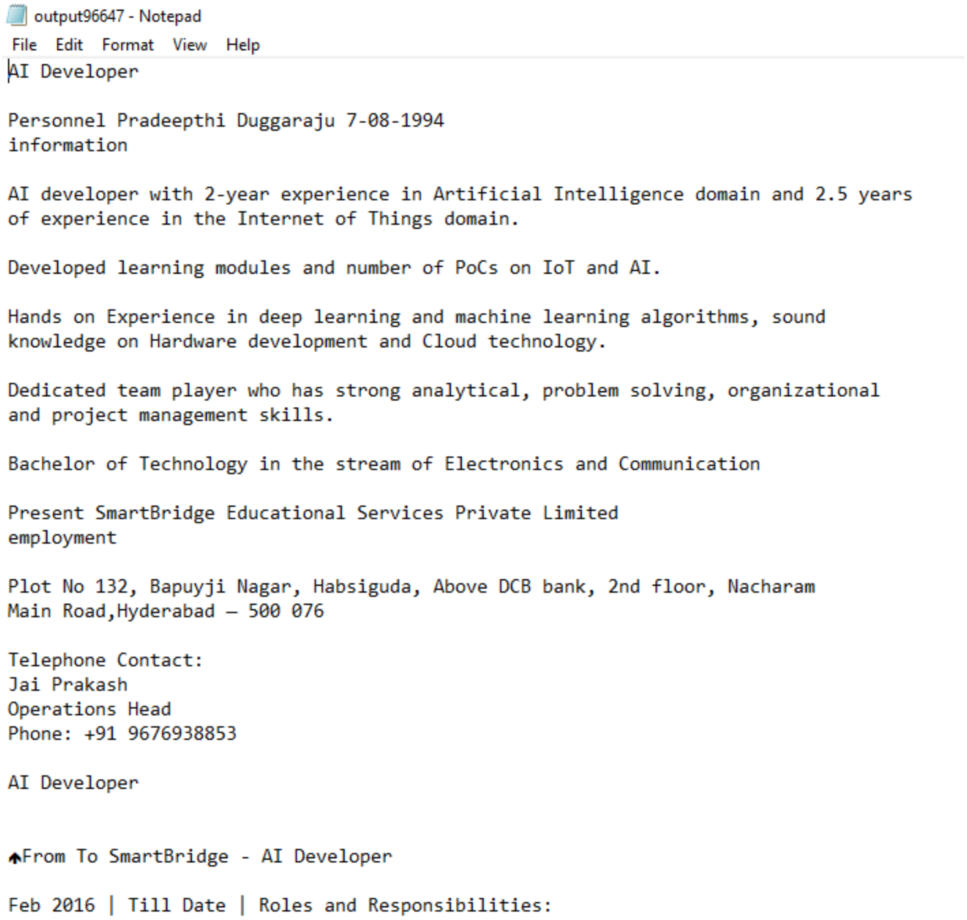
Optical character recognition or optical character reader is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image.

Upload PDF Here

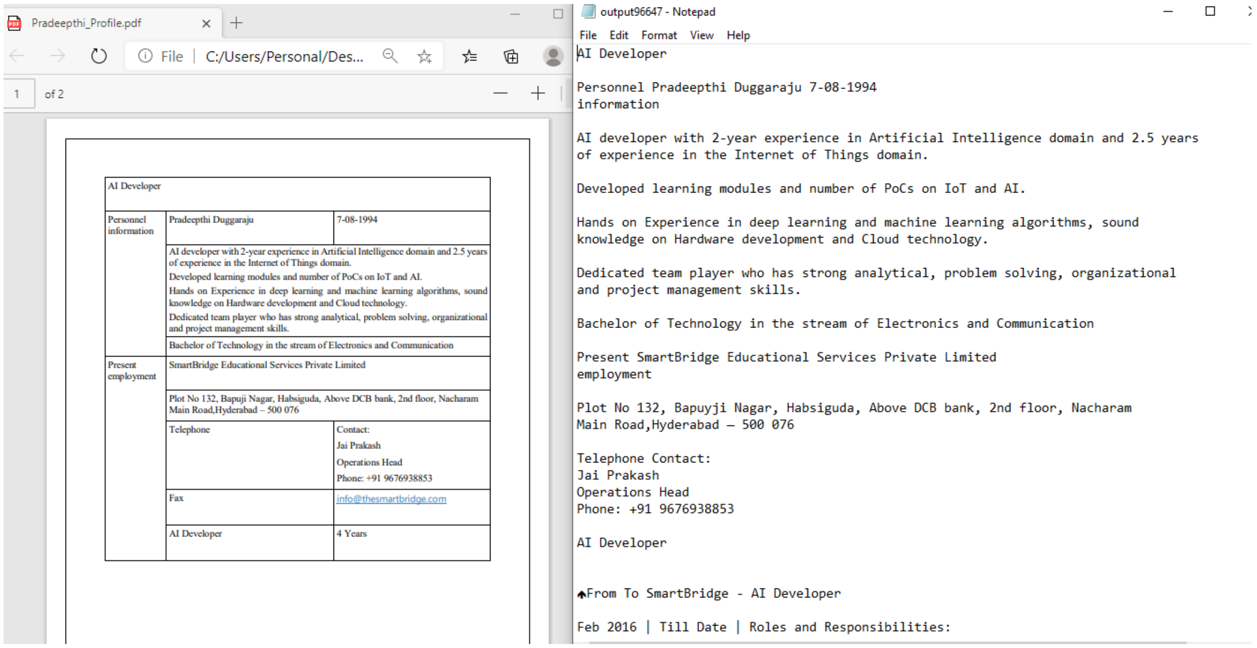
Converted file : C:\Users\Personal\Desktop\ocr - flask\outputs\output96647.txt



Name	Date modified	Type	Size
output4662	31-08-2020 08:25 PM	Text Document	1 KB
output37631	19-08-2020 07:43 PM	Text Document	1 KB
output64095	19-08-2020 05:56 PM	Text Document	0 KB
output65238	31-08-2020 08:18 PM	Text Document	0 KB
output96647	01-09-2020 02:08 AM	Text Document	2 KB

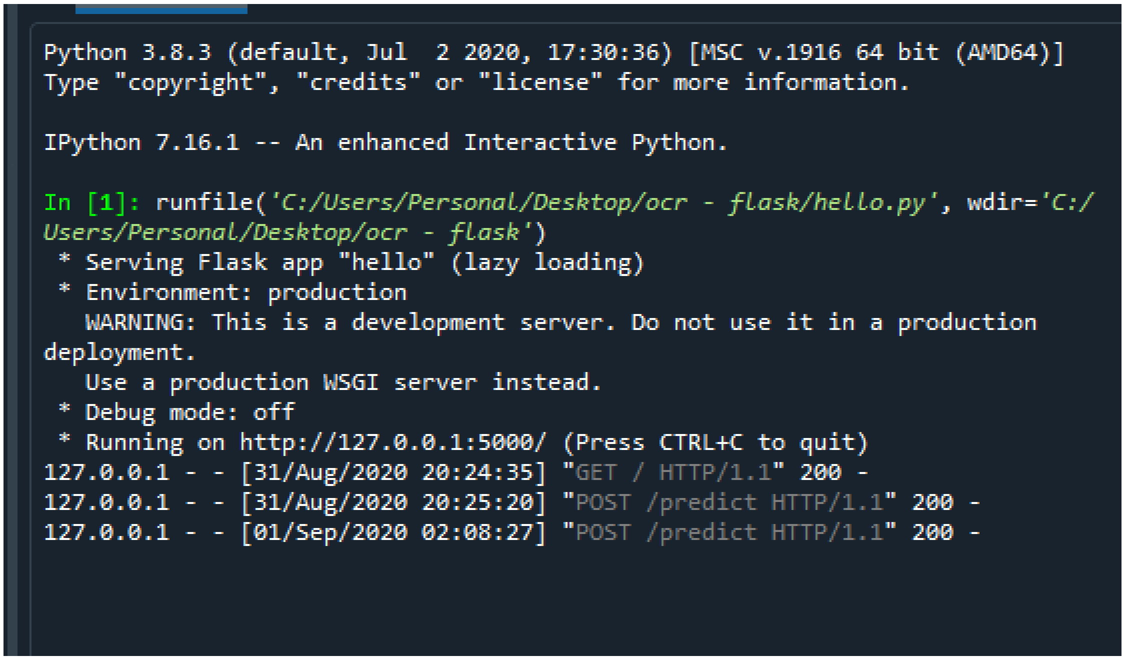


```
AI Developer

Personnel Pradeepthi Duggaraju 7-08-1994
information

AI developer with 2-year experience in Artificial Intelligence domain and 2.5 years
of experience in the Internet of Things domain.

Developed learning modules and number of PoCs on IoT and AI.

Hands on Experience in deep learning and machine learning algorithms, sound
knowledge on Hardware development and Cloud technology.

Dedicated team player who has strong analytical, problem solving, organizational
and project management skills.

Bachelor of Technology in the stream of Electronics and Communication

Present SmartBridge Educational Services Private Limited
employment

Plot No 132, Bapuyji Nagar, Habsiguda, Above DCB bank, 2nd floor, Nacharam
Main Road,Hyderabad – 500 076

Telephone Contact:
Jai Prakash
Operations Head
Phone: +91 9676938853

AI Developer


↑From To SmartBridge - AI Developer

Feb 2016 | Till Date | Roles and Responsibilities:
```



AI Developer		
Personnel information	Pradeepthi Duggaraju	7-08-1994
	AI developer with 2-year experience in Artificial Intelligence domain and 2.5 years of experience in the Internet of Things domain. Developed learning modules and number of PoCs on IoT and AI. Hands on Experience in deep learning and machine learning algorithms, sound knowledge on Hardware development and Cloud technology. Dedicated team player who has strong analytical, problem solving, organizational and project management skills.	
	Bachelor of Technology in the stream of Electronics and Communication	
Present employment	SmartBridge Educational Services Private Limited	
	Plot No 132, Bapuji Nagar, Habsiguda, Above DCB bank, 2nd floor, Nacharam Main Road,Hyderabad – 500 076	
	Telephone	Contact: Jai Prakash Operations Head Phone: +91 9676938853
	Fax	info@thesmartbridge.com
	AI Developer	4 Years

```
AI Developer

Personnel Pradeepthi Duggaraju 7-08-1994
information

AI developer with 2-year experience in Artificial Intelligence domain and 2.5 years
of experience in the Internet of Things domain.

Developed learning modules and number of PoCs on IoT and AI.

Hands on Experience in deep learning and machine learning algorithms, sound
knowledge on Hardware development and Cloud technology.

Dedicated team player who has strong analytical, problem solving, organizational
and project management skills.

Bachelor of Technology in the stream of Electronics and Communication

Present SmartBridge Educational Services Private Limited
employment

Plot No 132, Bapuyji Nagar, Habsiguda, Above DCB bank, 2nd floor, Nacharam
Main Road,Hyderabad – 500 076

Telephone Contact:
Jai Prakash
Operations Head
Phone: +91 9676938853

AI Developer


↑From To SmartBridge - AI Developer

Feb 2016 | Till Date | Roles and Responsibilities:
```



```
Python 3.8.3 (default, Jul  2 2020, 17:30:36) [MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 7.16.1 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/Personal/Desktop/ocr - flask/hello.py', wdir='C:/
Users/Personal/Desktop/ocr - flask')
 * Serving Flask app "hello" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production
deployment.
   Use a production WSGI server instead.
 * Debug mode: off
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
127.0.0.1 - - [31/Aug/2020 20:24:35] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [31/Aug/2020 20:25:20] "POST /predict HTTP/1.1" 200 -
127.0.0.1 - - [01/Sep/2020 02:08:27] "POST /predict HTTP/1.1" 200 -
```

3.APPLICATIONS:

1. Digitization of information can help your organisation move towards a paperless workflow.
2. It can help your organisation enable quicker and more convenient processes, enhance customer experience, increase employee satisfaction and reduce costs.
3. It can help you drive better compliance practices in your company while also providing better customer service and increasing the transparency in your organisation.
4. OCR can also be used for assisting blind and visually impaired people by scanning the printed data using OCR and reading it aloud using text reading technology.
5. OCR technology allows you to digitize:
    a. Forms - legal forms, government procedures, tax fillings, etc.
    b. ID cards - driver's license, passport, Aadhar card, etc.
    c. Legal documents - affidavits, tickets, bonds, etc.
    d. Bank statements - passbooks, account statements, cheques, etc.
    e. KYC information - ID cards, address proof.
    f. License plates - number plates in various languages.
    g. Shipping container numbers - container numbers written in any orientation.
    h. And much more…

4.CONCLUSION:

OCR technology provides fast, automated data capture which can save considerable time and labour costs of organisations. By this application the user can upload a pdf document, the document is analysed by Optical character recognition package to extract text from it. The extracted text is again saved in a text document in the local drive.

5.FUTURE SCOPE:

This project can be further extended for recognizing handwritten documents. This software can be further upgraded in which functionality can be added to train handwriting of a particular individual and then can be used to recognize documents written by that individual. Also, software can be trained to recognize handwriting of multiple individuals and also different fonts. There is also scope of increasing accuracy of the recognizer so that no manual watch should be needed on the software other than inputting the data. All the data recognized by the OCR can be formulated and entered into a database, which makes it more secure, redundant and easily accessible and useful. This can make use to convert normal images into database data. Normal text documents data can be easily migrated into a database. For Business purpose all the resumes, letters, complaints sent to a company can be directly entered into a database without the huge load and necessity for data entry jobs. This can decrease the money spent on Manuka labour and time. Thus, software can be atomized to a higher level.

~~THANK YOU~~